WEB CONTENT MANAGEMENT

Government Web sites contain records that document public transactions just like paper records and, as a result, a Web site must be retained like any other record. Because of the volatile nature of Web sites, however, Web record retention has remained a challenge for archivists and records managers across the country. Static sites are uncommon, especially in government, where policies, procedures, and public notifications posted on Web pages change frequently. The Library of Virginia's approach to the preservation of Web content, context, and structure combines the need for record retention and disposition policies with the desire to preserve a Web site's format (look and feel) for historical and reference purposes.

An agency's Web records are composed of all publicly available information on the agency's Internet Web site, as well as those records on their extranet or intranet. Each of these sites contains records that need to be managed.

- *Intranet*. An *intranet* is an internal Internet site that is only accessible to persons within an organization.
- *Extranet*. An *extranet* is an intranet site that is accessible only to selected individuals outside an organization.

In order to document a Web site properly, several factors must be taken into consideration and the following elements retained:

- *Content.* The actual HTML-encoded pages themselves and additional content files referenced therein or content created by end users interacting with the Web site. Analyze the pages of the Web site to determine which elements constitute public records.
- *Context*. Administrative and technical records necessary for or produced during the management of a Web site. Maintenance of these records provides a context for Web operations, which attests to the reliability, authenticity, and integrity of an agency or locality's Web site.
- *Structure*: For those Web sites that have been appraised as records, a site map indicating the arrangement of a Web site's content pages and software configuration files of content management systems. Structure also includes technical characteristics of the record (e.g., file format, data organization, page layout, hyperlinks, headers, footnotes).

Consider creating a file-naming protocol for Web pages to help ease management of the site. Having a common taxonomy is important, because it maintains consistency between Web,

database, and ECM systems. In addition, because Web sites are updated frequently by various people and groups, develop a method for designating and controlling versions. This practice will help ensure that Web site content remains trustworthy.

Content management systems (CMS) can be also used to manage the content of a Web site. A CMS consists of both a content management application (CMA) and a content delivery application (CDA). The CMA can relieve the Webmaster of many of the decisions and actions required to manage the creation, modification, and removal of content from a Web site. A CDA uses and compiles the content management information to update the Web site. A CMS can also be used to create audit trails associated with content that is created directly online.

Traditional records management techniques easily apply to relatively stable contextual and structural Web site records. Managing Web page content is much more complex. Web pages are fluid in nature, and when updates or redesign of Web site maps change the organization of Web content, it may be deemed necessary to set aside a new record-keeping copy of Web site content.

Like e-mail messages, Web sites should be maintained according to Record Retention and Disposition Schedules based on the content they contain rather than their format. A Web site may contain any number of record types, including but not limited to meeting minutes, annual reports, photographs, press releases, maps, organizational charts, policies and procedures, and mission statements, for example. Different formats exist even within Web pages, such as text, image, audio, or video files.

Just as individuals are responsible for maintaining other electronic records according to Record Retention and Disposition Schedules, so are they responsible for ensuring that the information they place on their Web sites is available elsewhere in another format. If a record is only available on a Web site, the Web site is considered the record copy, which must be retained according to the appropriate retention schedules based on the content it contains. Web pages, therefore, are not considered record copies as long as the information contained within them is retained elsewhere.

Library of Virginia Role in Web Site Archiving

The Library of Virginia collects, preserves, and provides access to all Web sites of Virginia's state government agencies in the executive, legislative, and judicial branches of government as described in this guideline. State government Web sites are selected for inclusion based on

intellectual content, research and educational use, and long-term benefit to the citizens of the Commonwealth.

A state government Web site is defined as the collection of all files identified by a state government domain for the purpose of providing publicly available information, affording access to government services, and/or conducting the state's business. Large, complex Web sites may span multiple servers and domains but are unified by Virginia government-related content.

All Web sites selected will be collected and preserved in the formats in which they were primarily distributed to the public. They will be made accessible from the Library of Virginia's catalog and/or Web site, as well as from the Archive-It Web site, a subscription service of the Internet Archive, a non-profit organization founded to build an Internet library offering permanent access for researchers, historians, and scholars to historical collections that exist in digital format.

The Library of Virginia will collect the following Web sites:

- All executive, legislative, and judicial branch state agencies, commissions, and boards as listed in *The Report of Secretary of the Commonwealth* (i.e., the Blue Book)
- All independent state agencies listed in *The Report of the Secretary of the Commonwealth* (i.e., the Blue Book)
- All statewide constitutional officers (e.g., Office of the Governor, Office of the Lt. Governor, and Office of the Attorney General), the Governor's Cabinet Secretaries, and the First Lady
- Gubernatorial initiatives and special projects (e.g., Smart Beginnings, Capitol Square renovations, the Springfield interchange, Jamestown 2007)

Web sites are collected on an established schedule, which is currently monthly for statewide constitutional officers, the Governor's Cabinet, gubernatorial initiatives, and the First Lady. All other Web sites will be crawled quarterly. In the occurrence of significant public safety and health incidents or other noteworthy events, the Library of Virginia may, at its discretion, alter the crawling frequency of state agency Web sites.

The Library will not collect:

- Public or private college and university Web sites. Due to the size and the access restrictions of these Web resources, and because of existing state Archives practice, these sites will not be archived.
- Non-state agency Web sites. In general, captured Web sites will contain only state government information. Non-state government Web sites may be considered for capture if they contain significant state government information and assist in the formation of government policy.

Technical limitations

The Library of Virginia has partnered with the Internet Archive to collect, preserve, and provide access to the Library's Web archive collections to the best of its abilities via the Archive-It service. Web content is harvested using the Heritrix Web crawler and archived content is indexed and searchable via the Internet Archive's Wayback Machine.

As a general rule, simple, static Web pages are the easiest to archive. Limitations to capturing and playing back archival Web content are as follows:

- When a dynamic page contains forms, JavaScript, images, streaming media, or other elements that require interaction with the originating host, the archived pages might not contain the original site's functionality.
- Database-driven Web sites can be very difficult to harvest. For example, if you need to fill in a form to get access to the content, such as with a search box, the harvester typically cannot retrieve the content.
- JavaScript elements often are hard to archive and even harder to display in the Wayback Machine, especially if they generate relative links (links that do not contain the full address of the linked page).
- Web site owners can specify files or directories to be excluded from a crawl, and can even create specific rules for different automated crawlers. All of this information is contained in a file called robots.txt. The Archive-It tool respects robots.txt. exclusion headers. The Library will make every effort to contact site owners to be sure that they allow the Archive-It crawler to have appropriate access to their site.
- Password-protected sites cannot be accessed by the crawler and therefore will not be archived.

- Links to sites that are not in the same domain as a URL identified for archiving will not be captured. For example, if the Secretary of Public Safety site has a link to the Red Cross, the Red Cross's site will not be captured. However, embedded files are crawled regardless of whether or not they come from an offsite host.

The Archive-It Web crawler does not apply to locality Web sites.